Trustworthy AI: Industry-Guided Tooling of the Methods

Zakaria Chihani
zakaria.chihani@cea.fr
Université Paris-Saclay, CEA, List
Palaiseau, France

ABSTRACT

The need to assess and validate the trustworthiness of AI (robustness, transparency, safety, security, *etc.,*) has been the subject of considerable academic work for some time now. A natural evolution of such research efforts is to have a tangible impact in the industrial sector and in the upcoming standards. To this end, theoretical feasibility of algorithmic methods is not enough: one needs to put these methods inside usable tools that can scale to real-world problems. Evidently, this need has not gone unnoticed either and several teams are actively working on maturing their tools further and further in a constant race with a very rapidly moving field. While fundamental research is a paramount bedrock, in the present communication, we want to focus on how far we have come in satisfying the goal of seeing AI safely permeating our future. To this end, we will give a brief overview of recent collaborations with industrial actors in an effort to give the reader a wider notion of trustworthiness, one that may come into play on their own use-cases.

KEYWORDS

Verification, Validation, Test, Explainability, Trustworthiness, Artificial Intelligence, Neural Networks, Formal Methods

ACM Reference Format:
Zakaria Chihani. 2024. Trustworthy AI: Industry-Guided Tooling of the Methods. In *Conference on AI Engineering Software Engineering for AI (CAIN 2024), April 14–15, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3644815.3644970

1 INTRODUCTION

While still debating some questions (such as the likelihood of achieving Artificial General Intelligence), the AI community in particular and most of the related stakeholders in general seem to be more or less convinced that AI winters are a thing of the past and that the current summer will never end. Indeed, this rapidly evolving field, especially through the recent Deep Learning advances, is too good at particular useful tasks to simply disappear. There is little doubt, for example, that neural networks (NN) are poised to permeate a growing number of everyday applications, including sensitive software where trust is paramount.

But as these artifacts move from a fad status to more stably ubiquitous components, their deepening interweaving with different aspects of society deserves a special attention to the development of methods and tools for an adequate characterization of AI trustworthiness. This colossal quest is made difficult by the intrinsic opacity of NN and their increasing size, making any method that can bring us closer to trustworthiness a precious commodity [2]. By presenting this work, we give concrete examples of how our tools were able to help our industrial partners to characterize certain desired properties of their use-cases. To fit in a an extended abstract, we selected two instances of the many collaborations we carried out with industrial partners, one in the context of testing, with Renault, and one for contextualised out-of-distribution detection, with Thales, which we briefly present in the following section.

2 METAMORPHIC TESTING

For this work [4], we relied on AIMOS (AI Metamorphism Observing Software), a CEA-developed software built on the widely used Metamorphic testing paradigm [1]. The goal was to guide the development of models by incorporating, in the selection process, early validation stages.

Renault's use-case came from a crucial step in the factory production of Renault cars is the control of the conformity of welds of rear axles. This control is realized by the analysis of the image of the weld by an algorithm which has been trained on weld images labelled by a professional operator. Renault sought ways of validating the quality of such algorithms, and their robustness to degradations that stem from the environment (described in Renault's Operational Design Domain) in which it is applied (different lighting conditions, different focus, *etc.,*)

By using AIMOS[1], we helped Renault compare several of their ML-based models in the process of selecting those which best fit their requirements (illustration in figure 1).

3 CODE: CONTEXTUALISED OUT-OF-DISTRIBUTION DETECTION

A fundamental need of software safety, not just for Thales, is arguably to be able to define a boundary between expected operational domain and the situation where the software should not be used, or at the very least raise an alarm. This separation is arduous to define for machine learning programs, especially for visual classifiers, because of the inability to formally or semi-formally specify such high-dimensional data (*e.g.,* what is a pedestrian?).

Several tools attempt to tackle this task. What CODE [5] provides is a more generalized method that comes with a built-in interpretability features that is provided by its underlying PARTICUL [6] method. The work carried out in collaboration with Thales aimed at evaluating CODE on a use-case relevant to their

CAIN 2024, April 14–15, 2024, Lisbon, Portugal
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0591-5/24/04
https://doi.org/10.1145/3644815.3644970

[1] Part of the supported tools of the CAISAR open-source platformhttps://www.caisar-platform.com/

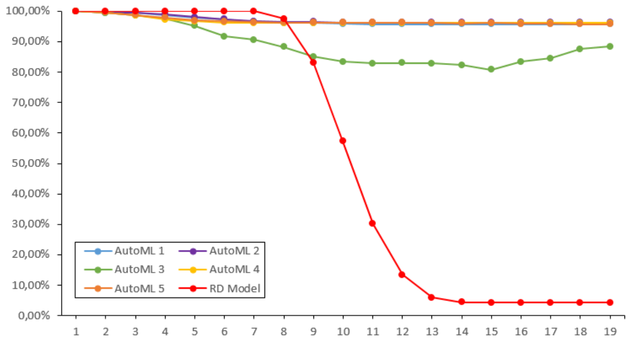

Figure 1: AIMOS on a Renault's welding belt: Stability of the models on the welds for the blur property with kernel size ranging from 1 to 20. Value points are linked *only* for aesthetic consideration to improve readability.

needs, and compare CEA's tool to state-of-the-art alternatives. The results are detailed in the co-written paper and briefly illustrated in figure 2.

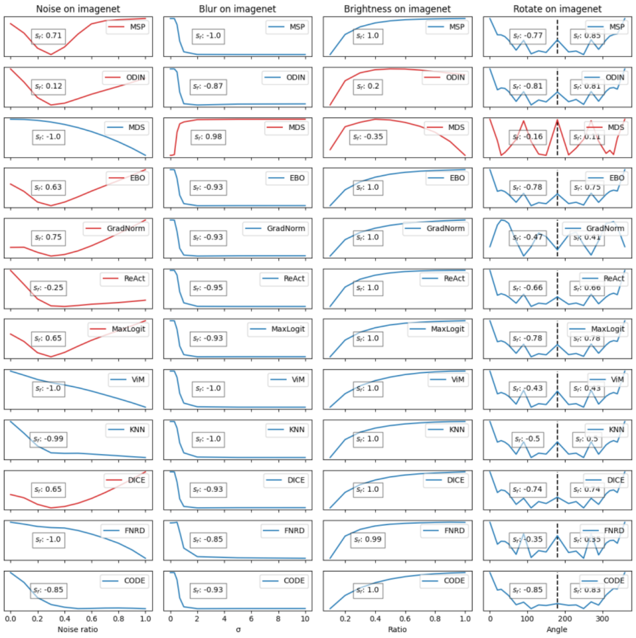

Figure 2: Evolution of the average confidence score v. magnitude α of the perturbation on the ImageNet dataset [3]. Curves in red indicate anomalous behaviours. Since all methods have different calibration values, we omit the units on the y-axis, focusing on the general evolution of the average confidence score over the perturbed dataset. Best viewed in colour.

4 CONCLUSION

Through these selected examples, we hope to bring a message forward to the industrial actors, seeking to increase trust in their models. It is possible to wait for the tools to mature on their own, through pure academic endeavour. However, what our experience shows, other than the results themselves, is that the collaboration offered to us by industry, through their real-world use-cases, their exigent requirements and their fast-paced development, can be a very powerful fuel. Rather than waiting for the train to move, we welcome your help in steering it.

SHORT BIO

In 2017, Dr. Zakaria Chihani founded the AI Trustworthiness division in the Software Safety and Security Lab (LSL) of the French Atomic Energy Commission (CEA). He has lead that division ever since and is now Deputy-Head of LSL. With tools and methods developed for explainability, testing and verification of AI, as well as a modular and extensible open-source platform for the characterization of AI Safety, his team is able to assist several industrial partners, such as Technip Energies or Renault, while striving to encourage academic effervescence around the topic through numerous talks, courses and event organizing such as WAISE and ForMaL. He helped shape the proposal for Confiance.ai, the 45m€ French project, with 13 founding partners and 32 associate partners. With well-established industrial actors such as AirLiquide, Airbus, Naval Group and Atos, this project was a very enriching experience for LSL's AI Trustworthiness team, and helped mature their tools and bring them ever-closer to full industrial applicability.

ACKNOWLEDGMENTS

Work partly funded under the Horizon Europe TRUMPET project grant no. 101070038 and the European Defence Fund AINCEPTION project grant no. 101103385. It was also supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute within the Confiance.ai Program (www.confiance.ai).

REFERENCES

[1] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, TH Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–27.

[2] Zakaria Chihani. 2021. Formal Methods for AI: Lessons from the past, promises of the future. (2021).

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Augustin Lemesle, Aymeric Varasse, Zakaria Chihani, and Dominique Tachet. 2023. AIMOS: Metamorphic Testing of AI - An Industrial Application. In *Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops: ASSURE, DECSoS, SASSUR, SENSEI, SRToITS, and WAISE, Toulouse, France, September 19, 2023, Proceedings* (Toulouse, France). Springer-Verlag, Berlin, Heidelberg, 328–340. https://doi.org/10.1007/978-3-031-40953-0_27

[5] Romain Xu-Darme, Julien Girard-Satabin, Darryl Hond, Gabriele Incorvaia, and Zakaria Chihani. 2023. Contextualised Out-of-Distribution Detection Using Pattern Identification. In *Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops*, Jérémie Guiochet, Stefano Tonetta, Erwin Schoitsch, Matthieu Roy, and Friedemann Bitsch (Eds.). Springer Nature Switzerland, Cham, 423–435.

[6] Romain Xu-Darme, Georges Quénot, Zakaria Chihani, and Marie-Christine Rousset. 2023. PARTICUL: Part Identification with Confidence Measure Using Unsupervised Learning. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, Jean-Jacques Rousseau and Bill Kapralos (Eds.). Springer Nature Switzerland, Cham, 173–187.